University of Southern Denmark
IMADA
Arthur Zimek
Jonatan Møller Gøttcke, Jonas Herskind Sejr

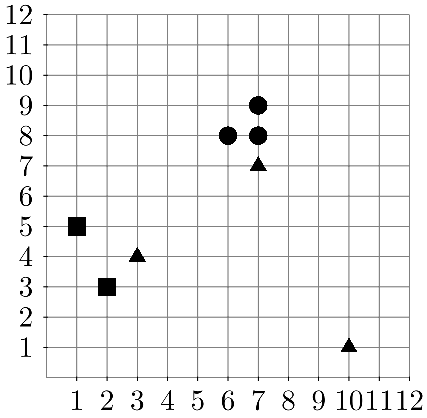
DM566/DM868/DM870: Data Mining and Machine Learning

Spring term 2019

Exercise 7: k-means, Evaluation of Classifiers

Exercise 7-1 k-means, choice of k, and compactness

Given the following data set with 8 objects (in $\mathbb{R}^2$) as in the lecture:



Compute a complete partitioning of the data set into $k = 3$ clusters using the basic k-means algorithm (due to Forgy and Lloyd). The initial assignment of objects to clusters is given using the triangle, square, and circle markers.

Objects x are assigned to the cluster with the least increase in squared deviations $SSQ(x, c)$ where c is the cluster center.

$$SSQ(x, c) = \sum_{i=1}^{d} |x_i - c_i|^2$$

Start with computing the initial centroids, and draw the cluster assignments after each step and explain the step. Remember to use the least squares assignment!

Give the final quality of the clustering (TD^2). How does it compare with the solutions for $k = 2$ discussed in the lecture? Can we conclude on $k = 3$ or $k = 2$ being the better parameter choice on this data set?

Also compute solutions with $k = 4$, $k = 5$, starting from some random initial assignments of objects to clusters. What do you observe in terms of the TD^2 measure?

Exercise 7-2 Measure for Evaluation of Classifiers

Given a data set with known class labels ($f(o)$) of the objects. In order to evaluate the quality of a classifier h, each object is additionally classified using h. The results are given in the table (all three columns) below.

ID	$f(o)$	$h(o)$
O_1	A	A
O_2	B	A
O_3	A	C
O_4	C	C
O_5	C	B
O_6	B	B
O_7	A	A
O_8	A	A
O_9	A	A
O_{10}	B	C
O_{11}	B	A
O_{12}	C	A
O_{13}	C	C
O_{14}	C	C
O_{15}	B	B

- Rewrite the definitions for precision and recall given in the lecture by using TP, TN, FP, and FN.
- Using the table (all three columns) above, compute precision and recall for each class.
- To get a complete measure for the quality of the classification with respect to a single class, the F_1-measure (the harmonic mean of precision and recall) is commonly used. It is defined as follows:

  $$F_1(h, i) = \frac{2 \cdot \text{Recall}(h, i) \cdot \text{Precision}(h, i)}{\text{Recall}(h, i) + \text{Precision}(h, i)}$$

  Compute the F_1-measure for all classes.
- So far, the F_1-measure is only defined for classes and not yet useful to get an overview of the overall performance of the classifiers. To achieve such an overall assessment, one commonly takes the average over all classes using one of the following two approaches:
  - Micro Average F_1-Measure: The values of TP, FP and FN are added up over all classes. Then precision, recall and F_1-measure are computed using these sums.
  - Macro Average F_1-Measure: Precision and recall are computed for each class individually, afterwards the average precision and average recall are used to compute the F_1-measure.

  Compute the Micro- and Macro-Average F_1-measures for the example above. What do you observe?

Exercise 7-3 Procedures for Evaluation of Classifiers

Given a data set D with objects from classes A and B ($D = A \cup B$) where the class assignments are *random* (not related to the attribute values). Furthermore, let the two classes have the same size $|A| = |B|$.

- What *true error rate* is to be expected for an *optimal* (for this data set) classifier?
- What error rates are to be expected when training and evaluating an optimal classifier on the given dataset using a leave-one-out test?
- Remember that in Bootstrap we produce the training and test data by sampling with replacement. An object is with a probability of

  $$\left(1 - \frac{1}{n}\right)^n \approx 0.368$$

  not part of the n training objects, i.e. only about 63.2% of the objects are used for training. (Compare this to 10-fold cross validation, where 90% of the data are used for training.)

  This implies that the error estimation is pessimistic, as the training set has size n, but actually only contains $0.632 \cdot n$ *different* examples.

  To make up for this, when evaluating bootstrap it is a common practice to also include the apparent classification error (error on the training data) during evaluation:

  $$\text{error rate} = 0.632 \cdot \text{Error on test set} + 0.368 \cdot \text{Error on training set}$$

  This will be repeated multiple times (with different samples) and averaged.

  What error rates are to be expected when evaluating an optimal classifier on the given dataset using the 0.632 Bootstrap method? Interpret these results.